MANUAL OF SAMPLING AND STATISTICAL METHODS FOR FISHERIES BIOLOGY

PART 1. SAMPLING METHODS

by

J.A. GULLAND

Fisheries Laboratory
Lowestoft, England

Fascicule 2

Section 2. Theory of sampling

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS
ROME 1966

CONTENTS

SECTION 2. Theory of sampling 1

2.1 Introduction 1

Example 2.1.1 1
Example 2.1.2 3

2.2 Random sampling 3

2.2.1 Random numbers 4

Example 2.2.1.1 5
Example 2.2.1.2 6

2.3 Stratified sampling 6

Example 2.3.1 8
Example 2.3.2 9
Example 2.3.3 10

2.4 Subsampling, or two-stage sampling 10

Example 2.4.1 12
Example 2.4.2 13
Example 2.4.3 14
Example 2.4.4 16

© FAO 1966
Printed in Italy

SECTION 2. THEORY OF SAMPLING

2.1 Introduction

Every sampling system is used to obtain estimates of certain properties of the population being studied, and the sampling system will be judged by how good the estimates obtained are. Individual estimates may, by chance, fall very close to or differ greatly from the true value, and give a poor measure of the merits of the system. A good sampling system will, on occasions, give an estimate which is far from the true value, just as a poor system may, very occasionally, give an estimate very close to the true value. A system is better judged by the frequency distribution of the many estimates which are, or could be, obtained by repeated sampling. A good system would give a frequency distribution with small variance, and mean estimate about the same as the true value. The difference between the mean estimate and the true value is called the bias. (The term "bias" is also used for the process whereby the difference arises.) The extent of bias and variance of a sampling system are to a large extend independent — a system can give estimates having small variance, i.e. differ little among themselves, but with a large bias, so that all the estimates differ greatly from the true value. (A measuring board with the figures on the scale almost illegible would introduce some extra variance; a board with the scale displaced to one side would introduce a bias.)

Example 2.1.1

Catches of mixed species of fish are examined by two observers for the percentage of one species of *Leiognathus*. Observer A works rapidly but carelessly, missing several fish of the required species; observer B works much more slowly but more carefully. From a succession of samples their estimates of the percentage of *Leiognathus splendens* in the catch were

A.	4	4	3	5	4
	3	5	4	6	4
	6	3	4	3	5
	4	5	4	4	6
	5	3	5	4	5

B.	9	7	11	4	8
	4	10	8	9	12
	8	3	6	10	15
	11	12	7	13	11
	10	5	8	9	12

By calculating the means and variances of the above, distributions show (*a*) that the estimates obtained by A are more precise (have less variance) than those of B (0.89:9.03); (*b*) that if, from other data, it is known that the true percentage was 9.1, then A has a strong negative bias; (*c*) if it is known that A misses half the fish, a relatively unbiased and precise estimate can be obtained by doubling the estimates obtained by A (mean 864, bias [i.e. the difference between the estimated mean and the true mean] –0.46, variance 3.6).

Bias may arise from a poor method of analysis, but more often from a poor choice of samples, or from the method whereby the measurement or counts are made or the samples are obtained, e.g. if the fish are sorted when landed, and the samples are mainly taken from the larger categories, the average size will be overestimated — a positive bias in mean size — or if plankton hauls are made with a coarse meshed net the numbers of small diatoms will be underestimated — negative bias in mean numbers of small diatoms, but a positive bias in mean diatom size. Bias in some form often occurs, particularly when attempting to obtain samples typical of the true condition in the sea, whether of diatoms with a plankton net, or of fish with a trawl.

If the size of sample increased, or the data of two or more samples combined, then the bias will remain unaltered, but the variance will be reduced, roughly in inverse proportion to the size of sample, or number of samples taken. The latter will in turn be closely related to the amount of work, or cost, involved in the sampling program. Any degree of precision[1], i.e. a variance as small as desired, can in theory at least, be obtained by taking large enough samples. The aim of good sampling is therefore not so much to obtain a given level of precision (small variance) but to do so with the least cost. Bias however, cannot be reduced purely by increased sampling, nor can its presence often be detected by subsequent analysis of the data (cf. Example 2.1.1 where there is nothing in the data themselves to tell which of samples A or B are biased). Bias can normally only be detected and hence eliminated by careful examination of the whole sampling process from beginning to end. In most situations great care should be taken to ensure that all likely sources of bias have been eliminated. There are however some situations where the bias is easily measured and it is simpler to allow the bias to occur, and to remove it in subsequent analysis (for example, gill nets are highly selective in the size of fish they catch, and hence provide a biased sample of say, the mean length. However, this selectivity can be measured and corrected in the later analysis). In this situation, as in all others, the

[1] A useful distinction will be maintained here between precision and accuracy; this corresponds closely to the distinction between variance and bias (or rather their reciprocals). A precise figure will have little variance and will be given with many significant digits, but may depart rather widely from the time value. If a fish is truly 17.638 cm long, precise measures of its length would be 17.64 cm, or 18.32 cm, but the latter is grossly inaccurate. Accurate but less precise measures are 17.6 cm or 18 cm.

chances of bias occurring must be fully examined before sampling, and if bias is accepted, its extent must be carefully measured, independently of the sampling process.

Example 2.1.2

In the situation of Example 2.1.1, larger samples are obtained by combining five of the original samples. Taking the mean of the rows as the estimate for the larger sample shows:

(*a*) that the variances of both sets have been reduced (that of A from 0.89 to 0.52, that of B from 9.03 to 1.29);

(*b*) that the extent of the bias of A's samples is unaltered (mean is unaltered).

The considerations above (bias eliminated, or at least known and measured, and variance a minimum for a given amount of sampling) will determine the sampling method, but the amount of sampling will be determined by the precision required. Usually it is not possible to state exactly the degree of precision required, but two limits can usually be given. At the lower limit the variance is so great that the information given by the sample has no practical value — the size of sample must be increased, or the sampling procedure abandoned. The estimates obtained from a single sampling scheme are usually combined with other data, some perhaps from other sampling systems, most of which will have a greater or lesser variance. The variance of the final answer will depend on variance of all the constituent pieces of data, but chiefly on the variance of the least accurate parts — a chain is as strong as its weakest link. For example, the total catch of a fleet may be estimated from data of the average catch per landing times the total number of landings. If the number of landings is only known within $\pm$ 10 percent, then however good the information on the average catch per landing, the total quantity landed will only be known within $\pm$ 10 percent at best. Once a certain level of precision has been achieved in a single sampling scheme, further improvement will not improve the precision of the final result, and the effort (time, manpower, etc.) would be better employed in improving the precision of other data.

2.2 Random sampling

The basic concept in all sampling is the random sample. A sample of objects from a population is random if all the members of the population have an equal chance of appearing in the sample. It is very important to remember that this applies to all members of the population,

exceptional as well as typical members. For instance, the whiting landed by any one ship at Lowestoft will usually (and here it will be assumed always) have a smooth unimodal length composition, with the mode usually between 28 and 30 cm, but occasionally, say once in 30 times, as high as 35 cm. A single sample of whiting from one ship if taken at random must therefore occasionally (once in 30 times on an average) give a mode at or above 35 cm, but usually around 28 to 30 cm. If then a fishery biologist, relying on one sample, obtained a mode at 35 cm, this departure from the average of 29 cm does not necessarily indicate nonrandomness because such misfortune will occur one in 30 times. The answer is to take more samples — all of 3 samples will have modes over 35 cm only once in 27,000 times.

2.2.1 Random numbers

A useful and widely applicable method of obtaining a truly random sample is by use of random numbers, as described in most statistical books. The individuals in the population from which a sample is to be drawn are allotted numbers, and those to be sampled are determined by reference to a table of random numbers. For instance, if a sample of 5 has to be taken from a population of 100, and the first 5 random numbers are 3, 47, 43, 73, and 86, the individuals corresponding to those numbers will be sampled. If the number of individuals in the population is not exactly 100 (or 1,000 etc.) some of the random numbers occurring will not correspond to numbers of the population and will have to be discarded. This wastage can be reduced if two or more numbers are ascribed to each individual, provided each individual has the same, i.e. has an equal chance of occurring in the samples. Suppose, for example, 5 units have to be sampled from a population of 24; in this case to each individual 4 numbers would be allocated — the first unit having for instance numbers 01 to 04 etc., the 24th having 93-96, so that only the numbers 97-100 are not used. The individuals to be sampled, corresponding to the previous set of five random numbers, would then be numbers 1, 12, 11, 16 and 22 (if one of the random numbers is 97 or over, it is discarded and another one taken). Instead of choosing all the units in the sample individually from the table of random numbers, units may be taken at regular intervals, e.g. every fifth or hundredth unit, and only the first chosen by use of random numbers. In the first example, one twentieth of the population is to be sampled, so that the sampling interval would be 20. The first random number is 3, so the complete sample would be the units numbered 3, 23, 43, 63, 83. Such a system can be dangerous when there is any natural periodicity in the population corresponding to the interval of sampling. For example, sampling at a landing place to obtain total catch should not be done every 7 or 14 days when there are great systematic differences between landings on different days of the week.

Example 2.2.1.1

Fish are landed at a certain landing place throughout the year. The total landings are to be estimated by sampling the catches in 30 days during the year. Determine the days on which sampling is to be done by using random numbers:

(*a*) directly from sets of random numbers 000 to 999, by numbering the days 1 to 365;

(*b*) by giving each day 2 numbers, 1 and 2 to 729 and 730;

(*c*) by giving each day 27 numbers, 1-27 up to 9829-9855, using random numbers 0000 to 9999;

(*d*) by sampling every 12 days, the starting day being chosen at random from days 1-12 (some samples may have 31 days).

If random numbers, or a similar randomizing process are not used, then it is likely that all individuals in the population will not have equal chances of appearing in the sample. If there is any correlation between the quantity being measured and probability of appearing in the sample, the result may be biased, perhaps strongly. For instance at a busy fish market it is often convenient when sampling the landings of a particular boat to work on the fish landed first. These will tend to be the fish caught last, and therefore the freshest. Such a sample would be heavily biased if used to estimate the freshness of the fish landed. The fish caught last will not however be very likely to differ in size from the other fish, so sampling the fish landed first may give unbiased estimates of the average size. The absence of bias must not be assumed too readily, and the possibility of bias arising must be carefully investigated. In the example above bias could arise if ships tended to do some fishing near their home port at the end of the trip, and if the size of fish on these grounds is different from the average. These and other sources of possible bias can only be detected and eliminated if there is a full knowledge and understanding of the fishery — how the fish are caught, how they are handled on board, and the arrangements of the fish market.

The precision of estimates obtained by true random sampling is readily determined. If a population is being sampled for some characteristic (e.g. vertebral count), which in the population has mean M, variance S^2, and a sample of n independent random individuals is taken with values $x_i \ldots x_n$, the resulting estimate of the mean value in the population is

$$\bar{x} = \frac{1}{n} \Sigma x_i \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.1)$$

and mean of $\bar{x} = M$ (the estimate is unbiased) and variance of $\bar{x}$ (or more shortly var $\bar{x}$) $= \frac{1}{n} S^2$, assuming that N, the total number in the population, is large compared with n.

Otherwise the formula for the variance becomes

$$\text{Var } \bar{x} = \frac{1}{n}\left(1 - \frac{n}{N}\right) S^2$$

Example 2.2.1.2

(*a*) Assuming the mean and variance of the data in Example 1.2.1 are close to the population values, calculate the variance in the estimated mean length from samples of 5, 20, and 100 fish;

(*b*) using random numbers, or otherwise, draw 20 random samples of 5 fish from the 449 fish in Example 1.2.1. Calculate the mean length of each of these samples; calculate the variance of these 20 values, and compare this with the expected variance as calculated in (*a*). (Note that the variance as calculated from a set of numbers as small as 20 is subject to some variability.);

(*c*) if it is required to estimate the mean length of North Sea cod to within $\pm$ 5 cm, how big a random sample is it necessary to take? (This requires that twice the standard deviation of the estimated mean length should be equal to 5.)

2.3 Stratified sampling

When sampling a heterogeneous population the precision achieved can be increased — sometimes very greatly — and the risk of bias reduced by dividing the population into sections, each relatively homogeneous, and sampling each section (or stratum) separately. Each stratum is then sampled independently, and estimates obtained for each. These can then be combined to give the estimate for the whole population. The variance of this estimate will also be obtained by combining the variances of the estimates within the individual strata. Since the within-strata variances will tend to be small — the strata being relatively homogeneous, so that the variance within strata is less, and possibly much less than the variance in the population as a whole — the variance of the final, combined estimate will also be small.

In mathematical terms suppose the population consists of N individuals, N_i in the i^{th} stratum where $N = \Sigma N_i$ and a sample of n_i is taken from the i^{th} stratum in which the values of the quantity to be estimated

(length of fish, weight caught etc.) are equal to y_{ij}, $j = 1 \ldots n_i$, the estimated mean value $\bar{y}_i$ in the stratum is

$$\bar{y}_i = \frac{1}{n_i} \sum_j y_{ij} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.2)$$

and an unbiased estimate of the mean value in the whole population is given as the weighted mean of the means of the individual strata, the weighting factor being the total numbers in each stratum

$$\text{i.e.} \quad \bar{y} = \frac{1}{N} \sum_i N_i \cdot \bar{y}_i$$

If the variance within the i^{th} strata is S_i^2

$$\text{var}\,(\bar{y}_i) = \frac{1}{n_i} S_i^2$$

and

$$\text{var}\,(\bar{y}) = \frac{1}{N^2} \sum \frac{N_i^2}{n_i} S_i^2 \quad \ldots\ldots\ldots\ldots\ldots \quad (2.3)$$

provided n_i is small compared with N_i. Otherwise the variances become

$$\text{var}\,(\bar{y}_i) = \frac{1}{n_i} \left(1 - \frac{n_i}{N_i}\right) S_i^2$$

$$\text{var}\,(\bar{y}) = \frac{1}{N^2} \sum N_i^2 \frac{1}{n_i} \left(1 - \frac{n_i}{N_i}\right) S_i^2$$

$$= \frac{1}{N^2} \sum N_i (N_i - n_i) \frac{1}{n_i} S_i^2$$

This variance may be compared with the variance of the estimate obtained by random sampling from the whole population, which is

$$\text{var}\,(\bar{y}') = \frac{1}{n} S^2$$

or

$$\text{var}\,(\bar{y}') = \frac{1}{n} \left(1 - \frac{n}{N}\right) S^2$$

if n is not small compared with N where S^2 is the variance in the population as a whole.

Example 2.3.1

The catch of a commercial trawler landing haddock at Aberdeen was sorted into four size categories forming the four strata (data from Pope, 1956). Samples of haddock from each category were measured, and the resulting data can be summarized as follows:

Category	N_i	n_i	Σy_{ij}	Σy_{ij}^2
Small	2 432	152	5 284	185 532
Small-medium	1 656	92	3 817	158 953
Medium	2 268	63	3 033	146 357
Large	665	35	2 027	118 169
TOTAL	7 021	342	14 161	609 011

where y = length of fish in cm.

From these figures, using as our estimate of S_i^2 the usual quantity

$$S_i^2 = \frac{1}{n_i - 1}\left[\Sigma y_{ij}^2 - \frac{1}{n_i}(\Sigma y_{ij})^2\right]$$

we have :

Category	$\bar{y}_i$	$N_i \bar{y}_i$	S_i^2	S_i^2/n_i	$N_i^2 S_i^2/n_i$
Small	34.763	84 544	12.21	0.0803	474 900
Small-medium	41.489	68 706	6.47	0.0703	192 800
Medium	48.143	109 188	5.48	0.0870	447 500
Large	57.914	38 513	22.85	0.6529	288 700
		300 951			1 403 900

$$\text{and hence } \bar{y} = \frac{300951}{7021} = 42.9$$

$$\text{and var } (\bar{y}) = \frac{1403900}{(7021)^2} = 0.0285$$

$$\text{and s.d. } (\bar{y}) = 0.17$$

The 95 percent confidence limits for the true mean length of the catch are therefore $42.9 \pm 2 \times 0.17$, i.e. $42.6 - 43.2$ cm.

The data can also be used to give a rough measure of the variance of the estimate obtained from a random sample of 342 from the whole catch. For this we shall take as the estimate of S^2, the variance of the population as a whole

$$341.\ s^2 = \Sigma y_{ij}^2 - \frac{1}{n}(\Sigma y_{ij})^2 = 609011 - \frac{1}{342} \times 14161^2$$

therefore $s^2 = 66.4$ (compared with the largest within strata variance of 22.85)

$$\text{var } (\bar{y}') = \frac{66.4}{432} = 0.197$$

$$\text{s.d. } (\bar{y}') = 0.44$$

The estimate of s^2 used is of course not entirely correct, as the sample used is far from being a true random sample, the medium fish being underrepresented. However it is sufficiently accurate to indicate the great reduction in variance in using stratified sampling, in this example a reduction to about one seventh of the unstratified variance, equivalent to an increase in sample size of seven times.

The benefit of stratified sampling will be increased by sampling the various strata in the best proportion. Those strata containing many individuals, or which are highly variable, need more sampling than small or uniform strata. The variance will be a minimum for a given total sample size, n, if

$$n_i \propto N_i \times S_i$$

or

$$\frac{n_i}{N_i} \propto S_i,$$

i.e., the proportion sampled within a strata is proportional to the variance in that strata. If n_i is not small compared with N_i this formula does not hold exactly, but will be close enough to give a good guide to the best allocation.

Example 2.3.2

In the Example 2.3.1 determine the best allocation to each strata of the total number of fish sampled (342), and using the values of S^2 calculate the variance of the estimated mean length for that allocation of samples.

Example 2.3.3

Along a certain coast the 100 places where fish are landed can be roughly graded into three classes according to the weight of fish landed. During one week the weights landed were:

Large landing places:	45	59	87	41	71	25	9	69	10	7
Medium:	17	13	19	26	1	8	27	11	12	26
	5	8	10	16	16	4	16	16	13	29
	14	25	29	27	20	25	2	7	3	12
Small:	2	6	7	0	1	2	1	5	4	7
	8	9	3	2	5	4	2	0	2	8
	5	3	8	9	8	9	1	6	5	3
	3	4	7	5	5	3	2	4	6	1
	6	2	5	1	0	3	8	0	4	3
	3	5	5	0	7	0	9	7	9	0

By calculating the variance within each class and in the population as a whole determine what would be the best method of estimating the total weekly catch along the whole coast, if the catch at only 20 places (one in five) can be determined (e.g. by visiting the landing places). What is the variance of this estimate, and how does it compare (*a*) with that of a simple random sample from the whole population, and (*b*) using stratified sampling, taking a sample of one fifth from each class?

2.4 Subsampling, or two-stage sampling

When the population being sampled is at all extensive or complex, the practical problems in taking a simple random sample are great, and the time taken for even a small sample may be large. The time required to obtain a sample of a given size may be greatly reduced by carrying out the sampling in two stages. First the complete population may be divided into a number of distinct primary units or subpopulations, and from these a sample is taken. From each of these sampled subpopulations a secondary sample, or subsample of individuals is taken. For instance, in estimating the total catch along a certain coastline, the basic unit may be taken as the landing of an individual boat. Taking landings at random along the whole coastline would mean an impossibly large amount of traveling; landings at a particular place on a particular day form a convenient primary unit. The procedure would then be to select (e.g. by random numbers) certain landing places on certain days, and at these selected places sample certain of the boats landing there.

Subsampling can of course be taken to more than two stages. If in the above example detailed examination, e.g. of size or maturity, was required, this might be done on a sample of a box of fish (or even a subsample from a box) landed by a certain boat at the landing place, giving three (or four) stages of sampling.

The disadvantage of subsampling is of course that individuals in the same primary unit are likely to be more nearly the same than individuals in the population as a whole. Thus, after examining one individual in the unit, e.g. weighing the catch of one boat landing at a particular place, examination of further individuals from that unit will tell one less about the characteristics of the whole population (e.g. the average catch per boat at all landing places) than examination of individuals from other primary units. This has to be balanced against the increased number of samples that can be taken in a given time by using two-stage sampling. In general terms, if individuals within a primary unit are very variable it is best to take many samples within a unit, with comparatively few primary units. Conversely if the variation within a unit is small, but there are considerable differences between units, then a large number of primary units should be sampled, with a small number of individuals sampled in each.

The method can be illustrated in mathematical terms: supposing for the sake of simplicity that the population can be split into K primary units, each of N individuals, and k primary units are sampled, a subsample of n individual being taken from each.

Then if M is the population mean, and M_i the mean for the i^{th} primary unit, we have as the estimate of the mean of any sampled primary unit:

$$m_i = \frac{1}{n} \sum_{j=2}^{n} x_{ij}$$

where x_{ij} is the value of the j^{th} individual in the i^{th} unit and the estimate of the population mean is

$$m = \frac{1}{k} \sum_{i=1}^{k} m_i \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.4)$$

Then the variance of m_i about M_i is $\frac{1}{n} \times S_w^2$, where S_w^2 is the variance of the individuals of the i^{th} primary unit about the unit mean. The variance of the estimated population mean will be made up of two parts — the variance of the estimated unit means about the true unit means, and the variance of the latter about the population mean; that is

$$\text{var}\,(m) = \frac{1}{k} S^2_B + \frac{1}{k} \cdot \frac{1}{n} \cdot S^2_w \quad \ldots\ldots\ldots \quad (2.5)$$

where S^2_B is the variance of the unit means about the population mean. An unbiased estimate of the variance of m is

$$s^2_m = \text{estimate of } S^2_m = \frac{1}{k}\,\frac{1}{(k-1)}\,\Sigma\,(m_i - m)^2$$

$$s^2_m = \frac{1}{k\,(k-1)}\left[\Sigma\, m_i^2 - \frac{1}{k}\,(\Sigma\, m_i)^2\right] \quad \ldots\ldots\ldots \quad (2.6)$$

Example 2.4.1 (from Pope, 1956)

A random sample of herring was drawn from the total number of landings in a week, and 50 herring from each selected landing were taken at random and measured. The following data were obtained:

Ship		1	2	3	4	5
Sum	$\Sigma_j x_{ij}$	1244.3	1324.2	1335.4	1299.7	1270.5
Sum of squares	$\Sigma_j x^2_{ij}$	31020.97	35127.08	35730.30	33900.99	32558.55

Estimate the mean length of herring in the week's landings, and its standard error. First we obtain the mean for each ship, equal to 24.9, 26.5, 26.7, 26.0 and 25.4. Therefore the estimates required are given by

$$m_2 = \frac{1}{5}\,(24.9 + \ldots + 25.4) = 25.9$$

$$s_m = \frac{1}{5 \times 4}\left[(24.9)^2 \ldots + (25.4)^2 - \frac{(24.9 + \ldots 25.4)^2}{5}\right] = 0.113$$

$$s_m = 0.34$$

The variances within and between primary units can also be estimated separately. Within any primary unit we have an estimate of S^2_w as

$$(S^2_w)\,i = \frac{1}{n-1}\,\sum_{j=1}^{n}\,(m_i - x_{ij})^2$$

$$= \frac{1}{n-1}\left(\Sigma\, x^2_{ij} - \frac{1}{n}\,(\Sigma\, x_{ij})^2\right)$$

These estimates from separate primary units can be combined to give as the best estimate

$$S^2_w = \frac{1}{k} \cdot \frac{1}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{n} (m_i - x_{ij})^2 \quad \ldots\ldots\ldots\ldots \quad (2.7)$$

From equations (2.5) and (2.6), the between units variance can be deduced from the equation

$$(k-1)\left(S^2_B + \frac{1}{n} S^2_w\right) = \sum_{i=1}^{k} (m_i - m)^2 \quad \ldots\ldots\ldots \quad (2.8)$$

and from the value of S^2_w given by equation (2.7)

Example 2.4.2

Calculate the within ship and between ship variance of herring lengths from the data of Example 2.4.1. We have as the estimate of within ship variance

$$5 \times 49 \times S^2_w = \left(31020.97 - \frac{1}{50}\ 1244.3^2\right) + \ldots + \ldots$$

$$\therefore 245\, S^2_w = 378.62$$

$$S^2_w = 1.545$$

$$\text{Also } S^2_B + \frac{1}{50} S^2_w = (24.9^2 + \ldots)^2$$

$$S^2_B = 0.56 - 0.03 = 0.53$$

In the calculations of Examples 2.4.1 and 2.4.2 it may be seen that the major contribution to S^2_m, the variance of the estimated mean length of all fish landed, comes from S^2_B, the between ship variance. Further, it follows from equation (2.5) that the effect of this on the variance of the mean can be reduced by increasing k, the number of primary units sampled, but not by increasing n, the number of individuals sampled in each primary unit. The time spent sampling the herring landings would therefore most probably be more efficiently used if the number of ships sampled was increased, at the expense of reducing the number of fish measured, e.g. 6 samples of $30 = 180$ fish, instead of 5 of $50 = 250$ fish. The best allocation of time will be determined by the time taken in the various stages of sampling, as well as on the variances

concerned. The total time spent can, to a first approximation, be split into three parts:

(*a*) the overhead time; the time spent in preparation, including time taken traveling to the sampling area from headquarters. This time is more or less fixed irrespective of the amount of sampling;

(*b*) the time between primary units — in the example the time spent moving from one ship to the next — which will be proportional to the number of primary units;

(*c*) the time within primary units — the time spent examining the individuals within each primary unit.

The total time spent will therefore be given by

$$t = t_o + k\,t_b + nk\,t_w \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.9)$$

when

t_o = overhead time
t_b = time moving from one primary unit to another
t_w = time spent examining one individual.

The best allocation (i.e. that giving the minimum variance) of sampling time, in terms of the number of individuals sampled in each primary unit, is given by

$$n = \frac{S_w}{S_b}\sqrt{\frac{t_b}{t_w}} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.10)$$

Example 2.4.3

Using the data of the previous examples, and assuming that 20 fish can be measured in a minute, and that the time taken to move from one ship to the next is 5 minutes, show that the least variance in the estimated mean lengths for a given amount of sampling is given by using secondary samples of about 17 fish.

So far it has been assumed that the primary units are all of the same size. When, as usually happens, they are of different sizes it is important that the correct weights are applied to each unit. Then equation (2.4) must be rewritten as

$$m = \frac{1}{N}\sum_{i=1}^{k} N_i\,m_i \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.11)$$

where N_i = number of individuals in the i^{th} primary unit, $N = \Sigma\,N_i$ = total number in all sampled primary units, or as

$$m = \frac{1}{N}\sum_{i=1}^{k}\left[N_i\,\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}\right] \quad \ldots\ldots\ldots\ldots\ldots \quad (2.12)$$

where n_i is the number of individuals sampled in the i^{th} primary unit, which will not necessarily be the same for all primary units. If n_i is taken so that the sampling ratio n_i/N_i is the same for all units, equal to p say, then (2.12) reduces to

$$m = \frac{1}{N} \sum_i \frac{1}{p} \sum_j x_{ij}$$

$$\therefore m = \frac{1}{n} \sum_i \sum_j x_{ij} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.13)$$

where n is the total number of individuals sampled. This, of course, is a most convenient form for computation. The formula for the variance (equation 2.5) has also to be rewritten, and becomes

$$\text{var } m = \frac{1}{N^2} \sum_{i=1}^{k} N^2_i \text{ var } m_i$$

$$\text{where var } m_i = \frac{1}{n_i} S^2_w + S^2_b$$

The formula (equation 2.10) for the best number to sample in each unit will also no longer be strictly applicable. The equation could be modified to give a formula determining precisely the best allocation to each primary unit sampled. However, this formula will be rather cumbersome and will need extra information on the variances within each primary unit (which may not be the same for each unit). The extra precision achieved may not be worth the effort involved, and a more reasonable method is to use equation (2.10), modified empirically by increasing the numbers sampled in the larger or more variable primary units.

When the object of the sampling is to measure some total quantity, e.g. the total weight landed of a certain species of fish, rather than some mean value, e.g. the average length of fish, the analysis of the results, as given in equations (2.11)-(2.13), must be modified. The total in the i^{th} unit sampled will be

$$X_i = \frac{N_i}{n_i} \sum_{i=1}^{n_i} x_i$$

where $\frac{N_i}{n_i}$ is the raising or weighting factor for the i^{th} primary unit, and is equal to the reciprocal of the proportion sampled. The total in the whole population is given by

$$X = \frac{N}{N_i} \cdot X_i$$

where N = total number of individuals in the population. If N is not known, as may well happen, then the proper raising factor N/N_i cannot be used, and the approximation K/k has to be used, where K is the total number of primary units and k is the number of units sampled (if the number of individuals in each primary unit is the same, the two raising factors will of course be equal). The use of the two raising factors in succession — from sampled individuals to whole primary unit, and from sampled primary unit to whole population — is most important. Serious bias can arise by using wrong weighting factors, if there are big differences in composition between primary units, especially if these are correlated with the number of individuals in the primary unit. Suppose for example we wish to estimate the total quantity of a certain species of fish living predominantly inshore, landed at a certain place. We may take as the primary unit the catch of a single boat, and sample from selected boats one box of fish. It is likely larger boats will work further offshore and have larger catches, and also have a smaller proportion of the inshore species in their catches. If the samples from these boats were only given the same weighting factor as those from the smaller inshore boats the proportion of the inshore species would be seriously overestimated.

Example 2.4.4

At a certain landing place 30 boats landed fish. One box of fish was sampled from each of 10 boats, and the weight of two species of fish determined, with the following results:

Boat numbr	1	2	3	4	5	6	7	8	9	10
Number of boxes landed	28	10	16	20	18	12	10	5	15	25
Weight of species A in 1 box (kg)	10	1	2	2	7	8	3	2	9	12
Weight of species B in 1 box (kg)	1	10	2	2	2	7	3	9	8	2

Calculate the total weight of each species landed (*a*) using information given above, (*b*) using the additional information that the total landings by all boats were 450 boxes. Compare the ratio of the 2 species in the total landings with the ratio in the 10 boxes actually sampled (one box equals 50 kg).

PRICE: $0.50 or 2s 6d

PM33271/9.66/E/I/3800